BEFORE THE EXECUTIVE COMMITTEE OF
THE EPA SCIENTIFIC ADVISORY BOARD

COMMENTS OF CHEMINOVA AGRO A/S ON THE MAY 31,2000, REPORT BY THE JOINT SAB/SAP SUBCOMMITTEE ON DATA FROM TESTING OF HUMAN SUBJECTS

JUNE 16, 2000

Cheminova Agro A/S is the registrant and principal manufacturer of several organophosphate insecticides. Cheminova recently completed and submitted to EPA an acute single-dose oral study of malathion, using human volunteer subjects. The study was conducted in accordance with the highest ethical standards at Inveresk Research, a world-class clinical laboratory that has performed dozens of similar studies on experimental human drugs for submission to the Food and Drug Administration. Within the last few years several other registrants have sponsored and submitted similar studies on other compounds. Cheminova greatly appreciates this opportunity to provide the Executive Committee of the Scientific Advisory Board (SAB) these comments on the *Report . . .On Data From Testing Of Human Subjects* (hereafter, "*Report*") prepared by a joint subcommittee of the EPA Scientific Advisory Board and the EPA FIFRA Scientific Advisory Panel (SAP). We also appreciate the significant effort of the Subcommittee and its support staff in working for a year and half on the content of the *Report*.

There is much in the *Report* that we agree with. The current draft is in many respects considerably improved over prior drafts that have become available. The *Report* now makes it clear that data from pesticide studies using human volunteers can provide a variety of benefits and are acceptable for many purposes if widely accepted ethical guidelines are followed. The *Report* also makes a series of useful recommendations about institutional changes and protocol guidance that EPA should adopt in order to clarify for future study planners what criteria should be used and what procedures should be applied in the design, review, conduct, and acceptance of new studies.

However, there is one major problem with the *Report*, and this is the focus of our comments today. Although not discussed in the executive summary or the transmittal letter, the body of the *Report* says that "the Subcommittee, in general, would not support human experimentation primarily to determine a NOAEL" because "[g]enerating such data pose[s] ethical concerns" (section 3.1, p. 10). Elsewhere the *Report* states that "generally, human dosing experiments are not appropriate if the primary intent of the study is to determine or revise a NOEL or NOAEL so as to eliminate the interspecies uncertainty factor" (section 3.2, p. 17). As we show below, however, the *Report* does not provide any support for this position.

This aspect of the *Report* is the only really important issue before the Executive Committee. It has become quite clear that the recently submitted human studies meet the ethical criteria that ordinarily govern the conduct of such studies. If accepted by the Agency, however, this newly created, intent-based criterion could provide an excuse for the Agency to refuse to use the recently submitted studies in setting RfDs. One of the major purposes of each of the recently submitted studies was to help in determining the proper NOELs and the proper uncertainty factors needed to set appropriate acute reference doses (RfDs) for the compounds in question. As the *Report* notes (p. 4), "when human data have been available and used it has generally raised the "safe dose." Opponents of the use of the studies in question have raised the "ethics" issue quite simply because they do not want to have the safe dose increased, and they are putting heavy pressure on Agency leadership to find a way to ignore the data, despite the Agency's long-standing practice of using human data when available in preference to animal studies.

Our comments are designed to explore the Subcommittee's apparent misunderstandings of the EPA regulatory system and the incorrect logic that underlie the *Report*'s disapproval of human testing for RfD purposes. The Executive Committee is well aware of how important it is that the SAB's recommendations are scientifically credible. We sincerely hope that the Committee will insist that the *Report* limit itself to the real scientific and ethical issues, and will direct that appropriate changes be made to the *Report* before approving it.

The Statements About Using Human Studies in RfD Setting Are Prospective in Nature

Despite our strong disagreement with the *Report*'s disapproving statements about use of studies to set RfDs, we note with appreciation that these statements are prospective, that is, they are aimed only at studies that might be conducted in the future. In sections 3.1 and 3.2 of the *Report*—where most of the purported reasons for the position are set forth—there is no mention of how to apply the principles to studies already performed and submitted. Instead, the discussion in section 3.2 ends with the statement that "the Agency needs to provide guidance both for its own policies and for parties contemplating the submission of human volunteer data" (p. 21).

The issue of EPA's use of studies that already have been conducted and submitted is taken up specifically in section 3.4.1 of the *Report*. The discussion starts (p. 29) by noting the controlling importance of the 1972 FIFRA amendment regarding the need for informed consent in human testing. The *Report* points out that "the 1972 statute must also be viewed as permitting use by the Agency of test data derived from human studies, when the law's strictures are met." It also says: "Useful data may, and often should, be used when they have been collected in compliance with any applicable law or regulation."

The section goes on to discuss (p. 30) whether "research results that have been obtained in a manner inconsistent with accepted ethical standards" may be used in decisionmaking. The Panel concluded (p. 31) that

> No algorithm can exist for making the decisions raised by this question. One can draw a temporal "bright line" benchmark, affirming that from a certain date, all research must meet certain ethical standards to be accepted by the Agency . . . But for prior research . . . there is an unavoidable need to rely on judgment. . . There will, of course, be transitional issues even if the Agency takes an unambiguously clear position for the future.

Accordingly, the *Report* does not condemn the use for RfD purposes of the human volunteer studies recently submitted to EPA, whatever the Agency may decide with regard to studies to be submitted in the future. This is especially true because the studies comply with the all traditional ethical criteria as set forth in the Helsinki Declaration, the federal Common Rule, and other similar codes requiring informed consent, use of institutional review boards, etc.

The *Report* Fails to Show Why Studies Using Human Subjects Should Not Be Used in Setting RfDs

The *Report* fails to provide any scientific or ethical basis for restricting the use of data that are gathered by generally accepted procedures from informed volunteer subjects. The *Report* puts forward several arguments, but none of these can survive examination. We discuss and respond to each of the arguments for the restriction on use of data in the discussion that follows.

Argument # 1: Human testing is not completely risk-free (p. 12).

Response: We agree, but in the studies that are at issue here—monitoring healthy adults for possible inhibition of blood cholinesterase after single doses of OP or carbamate pesticides—there is practically no risk of any significant adverse effect, given the nature and reversibility of the possible effect, the clinical setting and close medical monitoring, the single-dose regimen used for each subject, the stepwise nature of dosing, etc. The *Report* itself notes (p. 11) that studies of this type "are typically confined to low or moderate doses of limited duration and constructed as carefully as possible to avoid producing a serious effect, either acute or long term." As we pointed out in earlier comments, the Agency's 1998 *Neurotoxicity Risk Assessment Guidelines*, with which the SAB is quite familiar, stated that for these reasons ethically acceptable human studies could be performed to evaluate aspects of neurotoxicity such as cholinesterase inhibition. In the malathion study, no treatment-related effects were seen in any of the subjects at any of the five dose levels.

In addition, it is incorrect for the *Report* to suggest that to be of value, blood cholinesterase inhibition studies need to use doses high enough to yield a LOAEL, i.e., to cause an effect to be seen. All persons experienced in the field know that an inhibitory effect on blood cholinesterase would be seen in such a study if doses were sufficiently high. Thus, in such circumstances (a) there is no need to validate the test method by proving that effects will occur at a high enough level (indeed, would this not raise its own ethical questions?), and (b) a cholinesterase inhibition NOEL at the highest dose tested (HDT) is as valid as it would be if it were the next lower dose because an effect was seen at the HDT.

In any event, the possibility of some risk to subjects cannot be a reason for saying that data from an **already-completed** study, conducted in accordance with widely accepted ethical codes, can be used for one purpose but not another. In a completed study, any risk the study presented already has been incurred. Moreover, the recent studies all were designed and conducted at a time when there was every reason to think, based on EPA's written policies on neurotoxicity risk assessment and cholinesterase inhibition, that the Agency would accept and use them for setting reference doses. There is no basis whatsoever for **retrospectively** applying an "intent" test in order to protect their subjects from a "risk" that no longer can be experienced.

We also think there is no valid basis for **prospectively** applying an intent-based criterion to studies not yet conducted. The *Report* says it could be ethically permissible to conduct a human volunteer study if the sponsor is a professor seeking to advance scientific knowledge, or a pesticide manufacturer who will use the study results to improve the protocols for an upcoming animal study. Why, then, should it be ethically impermissible to conduct a study posing no greater risk, simply because the sponsor seeks to show that exposure to the residues of a pesticide—from use of a product to control disease vectors or make food more affordable—will not cause harm to humans? And why should it be impermissible to conduct a NOEL study on a pesticide with human volunteers, if FDA not only permits but actually requires the conduct of a healthy human volunteer study to observe what toxic effects are caused by higher and higher doses of a proposed new human drug?

Argument # 2: The NOEL for neurobehavioral effects might be lower than for other measures of toxicity (p. 13).

Response: Of course this is possible. For this reason, EPA has required studies of possible acute, subchronic, and developmental neurotoxicity in addition to studies regarding cholinesterase inhibition. The NOAELs from these studies for neurotoxic effects virtually always are higher than the blood cholinesterase inhibition NOELs. **Our position, simply, is that <u>if</u> EPA wishes to regulate on the basis of a blood cholinesterase inhibition endpoint, it makes more sense to use data from a proper <u>human</u> study than from a rat study of the same effect.** Of course another endpoint should be used if and when appropriate, e.g., if its use in combination with the appropriate safety factors results in a lower RfD than the RfD from the human volunteer study.

Argument # 3: Acute toxicity testing does not assess the potential for effects from long-term exposure (p. 13).

Response: This is certainly true, but this also is exactly why EPA requires a variety of longer-term studies and does not attempt to protect against possible chronic effects only by considering the results of acute toxicity studies. See the response to #2.

Argument # 4: "Dosing healthy adults provides extremely limited (if any) insights into the risks for the developing brain" (pp. 13-14). The Panel says this is the "most serious problem" with the human studies.

Response: This argument, like the two that precede it, is at odds with the fundamental aspects of OPP's system for assessing the possible risks of pesticides. The human studies in question are designed to determine the highest single test dose at which blood cholinesterase is not inhibited in healthy adults. We all understand that this kind of study will not evaluate the possibility of developmental effects in fetuses or infants. The need to match the potential endpoints of concern with studies designed to detect those endpoints is a fundamental principle of toxicology. It is why EPA has long required studies on developmental and reproductive effects and is now requiring developmental neurotoxicity studies on many compounds. The endpoints from developmental and reproductive effects studies are considered separately from acute toxicity results and can serve as the regulatory trigger when the facts dictate this, but ordinarily the NOAELs for developmental and reproductive effects are considerably higher than those for blood cholinesterase inhibition.

Argument # 5: The amount of exposure to pesticides that children experience may be greater than adults experience.

Response: This statement, which probably is true for some pesticides, has no relationship whatsoever to using tests with human subjects to set NOELs and RfDs. These studies do not measure the amount of exposure, whether to children or adults, but rather help evaluate what level of exposure is acceptable from a toxicity standpoint. This argument again ignores the Agency's entire approach to risk assessment, which requires both toxicity and exposure assessments.

We strongly urge the Executive Committee to recognize the mistaken nature of the *Report*'s criticisms discussed under ## 2-5 above and to direct that they be deleted from the *Report* before it is submitted to the Administrator.

Argument # 6: Studies need to have adequate statistical power (p. 15 and Appendices A and B).

Response: We agree, but this does not mean there is a problem with using the recently submitted studies. In December 1999 we submitted to the Panel detailed statistical information showing why the recent studies do in fact have adequate statistical power. As part of this showing, we explained why it is not necessary or appropriate, in the case of blood cholinesterase inhibition, to attempt to measure tiny differences, given the large amount of intra-individual and inter-individual variability in untreated humans. (For the benefit of the Committee we attach a copy of these comments.) We agree that a workshop to explore this issue is a good idea.

In the meantime, we think **Appendix B should be revised to remove the suggestion that for all purposes it is important to be able to measure very small percentage differences and thus that studies (in humans or animals) would need to use a much larger number of subjects than are called for by the Agency's current testing guidelines (all of which have been reviewed by the SAP and/or the SAB).** If taken literally, Appendix B could call into question the validity of those guidelines. It also could invalidate most of the Agency's animal toxicity database, since most animal studies use roughly the same number of subjects per dose group as the human studies at issue here.

Argument # 7: Certain problems are "associated with the use of NOAELs/LOAELs (e.g., design dependency, not an estimated value but the result of a test)" and thus "the Subcommittee does not believe human studies should be used to directly estimate these quantities."

Response: This seems to be an argument for doing away altogether with a regulatory system based on NOAELs in favor one using a different metric such as the ED_{10}. EPA has been considering such a change for a long time but has not yet made it.

We do not understand why this issue should be thought to bar the use of human data NOELs/NOAELs while allowing the use of comparable animal data NOELs/NOAELs. If EPA wishes to move to use of an ED_{10}-based approach, the data from the human studies could be used readily to generate the needed values. In the case of the malathion study, given the lack of response at the doses tested, this should result in a conclusion that the ED_{10} for blood cholinesterase is greater—probably considerably greater—than the highest dose tested. In short, this argument furnishes no reasoned basis for avoiding the use of human volunteer studies in setting RfDs.

Argument # 8: Finally, the *Report* sets up some straw men and then attacks them. It alludes to "ethical quandaries that arise when [human studies] are used simply to establish a NOAEL that lacks cogent scientific value and whose purpose can be interpreted as simply an argument for higher permissible exposure levels" (p. 11), although it never says what those "quandaries" are. Later the *Report* says that if the result of a human study "led to reduced use of pesticides, then the benefit of less pesticide in the environment could be realized . . . However, if the purpose is primarily to support the monetary gain of a company marketing a

product with no ability to rationalize the exposure in terms of general benefits to society, then the risk to individuals does not support this benefit" (p. 28).

Response: Both the statements above contain the same logical flaw. Of course, if a study—or a NOEL derived from it—"lacks cogent value" or cannot present any social benefit, then any risk to the human (or animal) subjects of the study would be difficult to justify. But to state what seems obvious, studies do not lack cogent value merely because they are designed to demonstrate what levels of pesticide exposure are safe for people to experience. To the contrary, the ability to use NOEL data from human studies in deciding what levels of human exposure are acceptable seems to us to have great value. We do not understand why anyone would think such data are not "cogent" (a term our dictionary defines as "believable" or "relevant"). Human risk is being evaluated—why would data from tests of rat blood be more believable or more relevant than parallel human data? What if the human study results showed that humans were considerably more susceptible than animals to the effects in question? (This possibility is one of the main reasons for Phase I experimental new drug trials in healthy human volunteers.)

Likewise, using the results of a study with human volunteer subjects to demonstrate that a valuable pesticide use can continue may be greatly beneficial to society at large, and this social benefit will be quite independent of the profits its registrant may realize from sales. Moreover, it ordinarily is regarded as quite appropriate to reward someone who makes available a product that society values because, e.g., the product improves the food supply or helps prevent disease. We do not see any ethical issue here.

We suggest that these passages in the *Report*—and other similar ones—should be deleted. If they are retained, most readers inevitably will conclude that the *Report*'s recommendations regarding using human studies to help set RfDs flowed from an underlying bias against pesticides, pesticide manufacturers, and/or our profit-oriented economy, rather than from any real concerns about scientific or ethical problems with human testing.

ATTACHMENT (CHEMINOVA COMMENTS, DECEMBER 23, 1999)

The Statistical Power Of A Human Study
To Detect Biologically Significant Differences
In Blood Cholinesterase Values

Comments Of Cheminova Agro A/S To The
SAB/SAP Joint Subcommittee On Data from Human Subjects

Prepared by Chris Wilkinson, Ph.D., Robert L. Sielken, Ph.D.,
Larry R. Holden, Ph.D., and Edward C. Gray,
Jellinek, Schwartz & Connolly, Inc.
December 23, 1999

These comments are submitted for the consideration of the U.S. EPA SAB/SAP Joint Subcommittee on Data from Human Subjects (the Subcommittee) by Cheminova Agro A/S. Cheminova is the sponsor of a recent study conducted at Inveresk Research (Inveresk) to determine whether malathion has effects on blood cholinesterase (ChE) levels in humans. Tests of this type have been performed during the past two years on a number of compounds; a great deal of information has been gathered, at a cost of several million dollars. We respectfully ask that the Subcommittee consider the following information in developing its final recommendations to EPA.

Overview

At its November 30, 1999 meeting, the Subcommittee indicated that tests using human subjects are never useful for the purpose of setting NOELs because, it was said, such tests always must be sufficiently powerful to detect differences as small as one percent between control and test groups, and that would require the use of thousands of test subjects. The conclusion that tests must be able to identify such small differences appeared to have been heavily influenced by (1) Dr. Herbert Needleman's position that detecting such small changes is important because some small changes could cause significant effects in the population at large, and his characterization of a study by Duffy et al. (1979) as showing that a single exposures to an organophosphate at a dose low enough to not cause immediately observable clinical symptoms could cause long-term adverse effects, and (2) Dr. Christopher Portier's position that, as a general design principle, tests must be able to detect very small changes in order to adequately protect individuals in the population at large whose sensitivity or period of exposure may be greater than the subjects of the test. As we explain below, Dr. Needleman incorrectly described the nature of study he referred to, and presented no other evidence that supports his

position as it relates to measuring inhibition of blood ChE, while Dr. Portier's position is inconsistent with EPA's current risk assessment policies and practices and thus does not evaluate the usefulness of either human or animal studies in the context of EPA's current practice.

We go on to show that such small changes in ChE inhibition are not of biological significance. To demonstrate this, we discuss the background variability in undosed humans in the malathion study. We also present the position of an authoritative international organization, the WHO/FAO Joint Meeting on Pesticide Residues, that statistically significant inhibition of 20% or more is regarded as biologically significant while inhibition at lesser levels must be reviewed on a case-by-case basis to determine whether it is of biological or regulatory significance.

Finally, we present information about the Inveresk malathion study. The study report, which soon will be submitted to EPA, shows that single oral doses of malathion did not have effects at any of the five doses tested. Attached to these comments is an analysis of data from that study that shows it has sufficient statistical power to provide meaningful and reliable scientific information at an appropriate level of biological relevance (e.g., power in excess of 0.9 to detect differences of 15% or more). We are providing this example to demonstrate that it would be incorrect for the Subcommittee to conclude that human studies of the type currently being conducted necessarily lack sufficient statistical power, although it is certainly proper for the Subcommittee to recommend that EPA give appropriate consideration to this aspect of test design.

Study purposes

At the outset, we should clarify the limited purposes for which the malathion study—and, we presume, other such studies—are intended. The study should not be considered a toxicity study and certainly was not intended to replace the vast number of animal studies on malathion that have focused on a variety of potential toxic endpoints. Under what EPA calls a weight-of-evidence approach to risk assessment, the Agency typically uses the lowest animal NOEL or NOAEL available and applies appropriate uncertainty factors to yield reference doses (RfDs) that then form the basis for EPA's risk assessments and regulatory positions. For the category of pesticides to which malathion belongs (organophosphorus insecticides, or OPs), inhibition of ChE in red blood cells (RBC) or blood plasma ordinarily is by far the most sensitive observable effect from acute exposure. Detectable inhibition of blood ChE typically occurs at dose levels well below those at which other, adverse effects (e.g., neurotoxic symptomology) occur. When this is found to be the case, EPA's practice (somewhat more conservative than that of most other regulatory agencies) has been to use the NOEL for inhibition of RBC or plasma ChE (not an adverse effect per se), with appropriate safety factors, as the basis for regulatory decisionmaking. Blood ChE inhibition is used as a biomarker of possible adverse cholinergic effects that may occur in other tissues or organs, usually only at considerably higher dose levels.

Cheminova thinks that if EPA proposes to base the allowable acute exposure level for a compound on a NOEL for blood ChE, it would be more appropriate to use human blood ChE data than animal blood ChE data from experimentally comparable animal studies, provided of course that the human data are scientifically valid and obtained from studies employing fully informed volunteer subjects. As late as 1998, EPA's repeatedly stated, official position has been to give preference to human data whenever it is available. Cheminova also thinks that the ability to compare the human and animal dose/response relationships and NOEL values for inhibition of ChE is extremely valuable for risk assessment purposes.

The Inveresk malathion study thus is intended to provide scientifically sound data indicating the level of a single oral dose of malathion causing no statistically significant inhibition of ChE in the red blood cells and plasma of humans. This information on the inhibition of human blood ChE could serve as the basis for establishing an acute RfD and/or help EPA determine whether it is appropriate to reduce the default (10X) uncertainty factor typically used to account for possible inter-species differences. In any event the information should lead to a better understanding of the circumstances under which this very widely used pesticide could be used with a low likelihood of causing significant acute toxicity. Thus, the availability of human data may help the Agency to conclude that malathion's availability for valuable public health and crop protection uses should continue.

The Subcommittee's November 30, 1999 conclusion that all studies must be able to detect very small differences is overbroad and lacks scientific support.

The Subcommittee's position that human studies always lack adequate statistical power for the purpose of setting NOAELs or RfDs was made after very little debate and without a broad inquiry into or consideration of the relevant facts. We think that such a categorical denial of the usefulness of studies in humans is inappropriate and overbroad because it ignores the differences in kinds and importance of effects that can be measured in studies in humans. In particular, we do not think the evidence and arguments that were presented to the Subcommittee support a ban on the use of human studies that evaluate whether a compound causes inhibition of blood ChE at particular dose levels.

> *No evidence was presented that long-term effects can result from exposures that do not cause acute clinical signs and symptoms.*

The Subcommittee position appears to have been influenced to a large extent by the written and oral presentations by Dr. Needleman. The entire evidentiary basis of his argument that the ability to detect small differences is important insofar as organophosphates are concerned was his description of a study of a group of people who, he said during the meeting, had "one exposure to organophosphates" and who a year later were found to have differences from controls in their EEGs. He argued that the study results show that "when you give a brain poison, particularly . . . organophosphates, and you say that you haven't produced an adverse

effect, you better be very sure that you haven't. And if that effect is very small, it requires large numbers of subjects." A little later in the meeting, when Dr. McConnell stated his understanding that the exposed subjects actually had shown actual clinical effects at the time of exposure, Dr. Needleman said, "No, you're wrong."

Actually, Dr. McConnell was right. The article in question is Duffy et al., Long-Term Effects Of An Organophosphate Upon The Human Electroencephalogram, *Toxicology and Applied Pharmacology* 47, 161-176 (1979). The study setting was extremely atypical: all the subjects were workers at an industrial facility where sarin nerve gas was handled on an ongoing basis. (Sarin is not an insecticide, but instead is a fluorinated OP compound designed as a potent neurotoxic warfare agent that could kill or incapacitate humans.) And the subjects all had shown clinical signs of poisoning at the time they were exposed to the gas, as the article states:

> Seventy-seven industrial workers with histories of accidental exposure to sarin were studied. This shall be referred to as the exposure (E) group. A subgroup of 41 E-group workers had histories of three or more exposures within the 6 years preceding the study. This subgroup shall be referred to as the maximum exposure (M) group . . . For the purpose of this study, an "exposure" is defined as (1) verified history of discrete exposure, (2) resultant clinical signs and symptoms consistent with exposure, and (3) reduction in erythrocyte cholinesterase to a level at least 25% below the individual's pre-exposure baseline. Every "exposure," as defined for the purpose of the study, was associated with equipment failure, operator error, or other industrial accident.

Thus, all the exposed subjects had indeed been subject to significant acute exposures at unknown levels that were high enough to have caused observable acute clinical effects at the time of exposure. The article does not say anything about the seriousness of the acute clinical effects, how much the ChE inhibition may have *exceeded* 25%, the level or duration of the accidental exposures of the various subjects, the maximum number of exposures suffered, or how long the subjects had worked at the facility (at least 6 years, apparently).

The Duffy et al. study clearly had nothing to do with setting NOELs for inhibition of blood ChE or with subtle differences in ChE levels. It is thus difficult to see why this study lends any support at all to the argument that detecting a 1-% difference in ChE inhibition in an acute exposure context is important or even useful.

In November 1999 the United Kingdom's Department of Health published a 250-page report, *Organophosphates*, that surveyed extensively the literature on the long-term effects of exposure to organophosphate insecticides (including Duffy et al). One of its conclusions is:

> No studies have examined the long-term effects of a single exposure to OPs insufficient to cause acute toxicity. However, the findings in individuals with prolonged and repeated low-dose exposures, and in those who have suffered recognized acute poisoning, together indicate than risk of serious health effects from such limited exposure must be small.

Dr. Needleman's arguments are unrelated to the ethics or usefulness of testing with human subjects.

Dr. Needleman's written presentation dated November 11, 1999 made a number of other points. For example, he argues that the nervous system has limited reparative capabilities; that the more sensitive the method used to measure effects the lower the NOAEL; that a small or rare effect can have important implications for populations; that children are more susceptible to neuroactive substances than adults; that detecting small or rare effects of exposure requires large numbers of subjects to avoid Type II risks; and that it's not known whether low doses of pesticides that pose no clinical symptoms may nonetheless cause subtle or latent effects. Some of these points are obviously true, while several are unproven and some are highly controversial. But the essential point is that *none* of them tends in any way to show that all studies of all effects in human volunteer subjects are unhelpful or necessarily lack adequate statistical power. Nor do any of these points address the question of why it always would be better to rely only on data from animal studies, which typically are no larger than human studies and have no more statistical power.

The Subcommittee's recommendations on statistical power of tests should be made in the context of EPA's current approach to risk assessment, not some proposed major restructuring of it.

The other major proponent of requiring studies to have adequate statistical power to detect very small changes was Dr. Portier. He also raised concerns about adequate statistical power to detect very small changes in ChE inhibition. His views undoubtedly are afforded great weight in this area because he is a biostatistician and the only statistician on the Subcommittee. But in his written comments to the Subcommittee, dated November 30, 1999, his call for the ability to detect very small changes is rooted not in either statistics or ethics, but in his desired approach to the risk assessment process.

EPA's current approach uses test data to estimate the dose response relationship (that is, how the mean response in that population changes with dose) in the population from which the test group was selected. It then uses an intraspecies uncertainty factor (and the additional FQPA children's safety factor, when appropriate) to account for human intraspecies variability. It does this whether or not human data are used; if animal data form the basis for regulation, an interspecies factor ordinarily is used as well.

Dr. Portier's written comments argue that a test's design should be such as to allow it to detect, with great statistical power, a quite low percent change such as 5% or even 1%. He says the reason for this is to "tak[e] into account sensitive individuals and possible effects of longer exposures in the environment as compared to the laboratory." In other words, he is saying, intraspecies (human) variability should be accounted for by increasing dramatically the size of groups used in toxicity testing, instead of (or in addition to) applying a 10X intraspecies uncertainty factor and, as appropriate, an additional 10X children's safety factor (EPA's current approach).[1] And he would account for uncertainties related to the duration of real-world exposure by increasing the size of groups used in short-term toxicity testing, instead of (or in addition to) EPA's current approach of requiring the conduct of a series of studies of appropriate dosing duration to establish toxicity endpoints for various exposure intervals. His approach to test size would apply equally to animal studies. In effect, Dr. Portier is suggesting that the entire approach to test design and risk assessment now used by EPA (and other agencies) should be dramatically changed. One might agree or disagree with some of the ideas he proposes, but it is clear that his paper's arguments about the need to detect 1 % differences are not aimed at the usefulness or appropriateness of human testing within OPP's current framework for assessing risk.

The Subcommittee should address statistical power of tests in the context of the current framework in which the tests could be used. The Subcommittee should assume that intraspecies differences will be adequately accounted for by EPA's use of appropriate uncertainty factors. The Subcommittee should not, either expressly or by implication, propose significant changes in the general approach to EPA risk assessment. Nor should it try to decide, for various kinds of potential adverse or non-adverse effects, what particular degrees of differences between test and control groups are of toxicological significance. That is not what the Subcommittee was convened to do, and that is not where the Subcommittee's expertise lies. The Subcommittee can make useful comments on the need for test power without exceeding its charge.

Statistical power in the context of blood cholinesterase measurements

The power of a study of a given size to detect a difference between dosed and placebo groups is a function of, among other things, the size of the difference that one needs to detect. We agree that very large groups are needed in order to be able to say reliably that a true 1% difference between tested and control groups was not missed in a test in which no effect was observed. We also agree that it is important to take needed steps to detect relatively small increases in the number of treatment-related effects that cause significant consequences in humans. (We do note, however, that this is as true of animal studies as it is of human studies.

---

[1] Similar suggestions for basic changes in the overall assessment to risk assessment, less openly stated, are found in comments by other Subcommittee members to the effect that testing in adults cannot be useful to protect children.

Most animal studies designed to measure the toxic effects of pesticides are of a size similar to or smaller than the current human studies.)

However, it should be emphasized that at least with respect to blood ChE levels, small differences are neither biologically relevant nor relevant to human risk assessment. As noted earlier, blood ChE inhibition is not itself an adverse effect, but an easily measured, sensitive biomarker that a recent SAP Subcommittee called an "imperfect mirror" of potential toxicity that may occur at other sites if the ChE level at those sites is sufficiently inhibited.

In addition, it is universally recognized that undosed, healthy individuals typically have wide variations in their blood ChE measurements over relatively short periods of time, and these variations do not appear to cause adverse effects. For instance, the blood ChE of each subject in the Inveresk malathion study was measured at sixteen different times over a period of about three weeks. Eleven male subjects received a placebo dose only; of them, the individual with the least downward deviation nonetheless had one RBC ChE measurement that was 12% below the mean of all sixteen measurements, and one person had a measured value that was 34% below the mean. The group means also vary considerably from interval to interval. For instance, the mean of the RBC ChE readings for all 11 placebos at 8 hours after dosing was more than 6% lower than the mean of the same group only 4 hours earlier. This kind of variability is not at all unusual and serves to place any discussion of the significance of 1% or even 5% changes in proper perspective.

Toxicologists differ somewhat in their approaches to the issue of what kinds of differences should be regarded as having biological relevance and regulatory significance. Of particular importance are the views of the Joint Meeting on Pesticide Residues (JMPR), an FAO/WHO organization that recommends maximum residue levels (MRLs, international pesticide tolerances) under the UN's Codex Alimentarius Commission system.[2] The JMPR's views on what levels of red blood cell ChE inhibition are of biological (and thus regulatory) significance are set forth in its *1998 Report*:

> Regulatory agencies have traditionally used various thresholds, such as 10% inhibition, 20% inhibition, or any statistically significant inhibition, in defining biologically significant depression of enzyme activity. The Meeting considered that statistically significant inhibition by 20% or more represents a

[2] The FQPA provides that Codex tolerances by default are governing for EPA tolerance-setting purposes. Section 408(b)(4) of the Federal Food, Drug, and Cosmetics Act, as amended in 1996 by the FQPA, says that EPA cannot establish a tolerance that differs from a Codex MRL without first proposing to do so and explaining the basis for the difference. The Codex MRLs also have an even more significant default status under the treaties and agreements administered by the World Trade Organization.

> clear toxicological effect and any decision to dismiss such findings should be justified. The Meeting also agreed that statistically significant inhibition of less than 20% or statistically insignificant inhibition above 20% indicate that a more detailed analysis of the data should be undertaken. The toxicological significance of these findings should be determined on a case-by-case basis. Considerations affecting such determinations include *inter alia* the shape or slope of the dose-response curve, assay variability, and correlation with clinical signs.

It is thus evident that there is no single magic number by which to gauge the biological significance of ChE inhibition. There is international consensus that a statistically significant decrease as high as 20% is of regulatory significance, and that, depending on the results of a case-by-case analysis, smaller variations that are statistically significant may or may not be regarded as having biological and regulatory significance. This obviously is a science policy matter upon which reasonable toxicologists and regulatory organizations may and do differ. The EPA Office of Pesticide Programs is now considering what criteria it should adopt in its deliberations on a draft Science Policy document on ChE inhibition issues. Most importantly for present purposes, this issue is not one that has been debated by the Subcommittee, nor is its resolution within the scope of the Subcommittee's responsibility.

Accordingly, if human studies like those recently conducted are sufficiently powerful when the detection difference considered biologically significant is in the range of 15% or 20%, it would be unscientific for the Subcommittee to conclude that human studies always lack enough power for use in establishing NOELs. The document that accompanies this one demonstrates that the malathion study's statistical power is more than adequate for such a purpose. Other studies of this type probably have similar power. Moreover, we think that the placebo data from a number of closely similar studies conducted at the same facility during the same period can be pooled to improve the power of the individual studies.

Thus, the Subcommittee cannot justifiably conclude that human studies are never sufficiently powerful to justify their use in establishing NOAELs, or that as a result they are somehow unethical no matter how low the risk and how well the informed consent and institutional review board aspects have been handled. Instead, the Subcommittee could conclude that studies that are manifestly inadequate with respect to statistical power should not be undertaken, and that such studies should not form the sole basis for regulation.

In this regard, it should be kept in mind that there are competing considerations with regard to study size. As the Subcommittee recognized, a small study may be powerful enough to identify an effect at a particular dose, even if the same sized study is not sufficiently powerful to demonstrate the lack of effect. Clearly, it would be ethically inappropriate to test more people than necessary at a level that causes an effect. What the Subcommittee apparently failed to note is that if an effect is not seen at a particular dose, a study of this kind can be

expanded sequentially to increase its power by adding further subjects at the same dose. This allows the risk of observing an effect in a large number of subjects to be minimized while also allowing a more robust examination of the initial indication that the compound does not cause the effect.

The Subcommittee also should consider that the various pesticide studies in animals and humans can and should be viewed as complementary. It is not just the results of a particular statistical power test applied to each individual data set that tell us how confident to be about the correctness of a regulatory decision. Within a study, a series of consistent results can lend confidence to the correctness of the overall conclusion, as can the consistent results of a set of various statistical analyses (e.g., analyses of individuals' variations from baseline over time and analyses of comparisons of group and control means at any given time). Different studies with consistent results likewise make regulators more confident of their readings of each study individually. One presumes that this is a major reason why EPA, FDA, and other regulatory agencies do not require animal studies to be larger than they are. Chronic and subchronic toxicity studies in dogs, for example, typically employ dose groups of no more than four animals per group.

Analyses demonstrate that studies like Cheminova's are able to detect important differences with adequate statistical power.

Attached is a document (Sielken and Holden, "The Substantial Power of Human-Testing Data to Contribute to the Dose-Response Characterization of Cholinesterase Inhibition in Humans: A Statistical Analysis") that presents and explains statistical power analyses of the Inveresk malathion human study RBC data. The document summary shows that the power of this study to detect variations in ChE of 15% by each of two completely different methods is greater than 90% (0.9):

> Human studies like the Malathion Human Cholinesterase Study initiated by Cheminova have substantial power to detect a dose that would cause cholinesterase inhibition in excess of 15%. For example, in the Cheminova study with an estimated standard deviation of 10% in a person's RBC cholinesterase inhibition, the one-sided, two-sample t-tests at a 5% significance level have a power of 92% at each of the highest doses to detect a 15% reduction in the mean cholinesterase level, and the trend test at a 5% significance level has a power of 96% to detect a 15% reduction in the mean cholinesterase levels by the highest dose.

The Inveresk malathion study has not yet been submitted to OPP; preparation of the final report is in progress. However, the cholinesterase values can be made available to the Subcommittee if that is desired.

.* * * *

Cheminova Agro A/S appreciates the opportunity to present these comments and hopes that the Subcommittee finds them helpful. The authors would be happy to provide the Subcommittee any further information upon request.

Attachment

The Substantial Power Of Human-Testing Data To Contribute To The Dose-Response Characterization Of Cholinesterase Inhibition In Humans: A Statistical Analysis

Robert L. Sielken Jr., Ph.D., and Larry R. Holden, Ph.D.
JSC Sielken, 3833 Texas Avenue, Suite 230, Bryan, TX 77802
Tel: 409-846-5175; Fax: 409-846-2671; E-mail: SielkenINC@aol.com

Executive Summary

Human studies like the Malathion Human Cholinesterase Study initiated by Cheminova have substantial power to detect a dose that would cause cholinesterase inhibition in excess of 15%. For example, in the Cheminova study with an estimated standard deviation of 10% in a person's RBC cholinesterase inhibition, the one-sided, two-sample t-tests at a 5% significance level have a power of 92% at each of the highest doses to detect a 15% reduction in the mean cholinesterase level, and the trend test at a 5% significance level has a power of 96% to detect a 15% reduction in the mean cholinesterase levels by the highest dose.

Human studies, like their animal bioassay counterparts, are designed to provide information on the dose-response relationship over the range of doses tested. Specifically, the primary statistical objective of these studies is to provide data on how the mean response in the test population changes as the dose changes. The characterization of this dose-response relationship can be used to identify no-observed-adverse-effect-levels (NOAELs) or risk-specific benchmark doses like estimated doses (EDs) or their lower bounds (LEDs) for the population from which the test subjects were randomly selected.

The information on how the mean response in the test population changes as the dose changes that is provided by a human experiment is combined with other information in the human risk characterization. As with animal bioassays, other information is used to characterize inter-individual variability within the test population and to extrapolate from the test population to other populations. For example, the uncertainty factor used to characterize intraspecies variability is not based solely on the data collected to characterize the dose-response relationship for the mean. Similarly, the uncertainty factor used to extrapolate from a test population to another population is not based solely on the data collected to characterize the

dose-response relationship for the mean. Thus, during the human risk characterization, the data collected from human testing to characterize how the mean response in the test population changes as the dose changes is supplemented with other information to identify individual differences in sensitivity, to identify potential sensitive subpopulations, and to extrapolate from adults to children. Furthermore, human testing to generate data on the dose-response relationship of the mean in humans for a specific health endpoint is usually supplemented with a broad database of information derived from animal testing including tests on a wide variety of potential health endpoints.

Power calculations are an important element of both experimental design and data interpretation. It is essential that power calculations accurately reflect the objective of the testing. Human studies like the Malathion Human Cholinesterase Study initiated by Cheminova have substantial power to do the job they were intended to do. That job is to characterize the dose-response relationship describing how the mean response in the test population changes as the dose changes in the range of doses tested. In such studies, the level of red blood cell (RBC) or plasma cholinesterase that is biologically significant as a biomarker for an adverse human health effect is generally believed to be of the order of 20% or higher. The size of the human study initiated by Cheminova has power in excess of 90% to detect these levels of cholinesterase inhibition.

Description Of Data

Data available for this analysis of power are the RBC cholinesterase (ChE) results for 38 male subjects in the Malathion Human Cholinesterase Study initiated by Cheminova. The subjects were distributed among six classes: 5 dose groups (0.5, 1.5, 5, 10, and 15 mg/kg) and one placebo group. There were 16 repeated ChE measurements per subject. Six of these were obtained prior to dosing and the remaining 10 obtained at different post treatment times. The average of the six pre-dosing values is used as a baseline level for each subject. Raw ChE values are then re-expressed as a percent decrease from baseline ChE. The specific number of subjects in each group is listed in Table 1.

Although the experimental design was organized in a repeated-measures format, the re-expression of raw ChE measurements as decreases from baseline removed most, if not all, of the intra-subject correlation. Thus, for the purposes of power calculation, the comparison between any two dose groups for each time point can assume a simple one-way analysis of variance model. Specifically, a formal test of a ChE decrease from placebo would use the t-statistic:

$$(1) \qquad t = (M_0 - M_d)/SE_{dif} \ ,$$

where M_0 is the sample mean for the placebo group, M_d is the sample mean for dose d, and SE_{diff} is the estimated standard error of this difference in sample means. This standard error is based on the usual one-way analysis of variance estimator:

$$(2) \qquad SE_{diff} = [\ S^2\ (1/n_0 + 1/n_d)\ ]^{½} \ ,$$

where n_0 and n_d are the number of subjects and S is usual one-way analysis of variance estimate of the standard deviation among subjects. That is,

$$(3) \qquad S^2 = \{S_i S_j (X_{ij} - M_i)^2\ \}/S_i(n_i-1) \ ,$$

where X_{ij} is the response of subject j in dose group i. The quantity $S_i(n_i-1)$ in the denominator is the 'degrees of freedom' (or simply df) and represents the amount of information about the inherent between-subject variation contained in S. Note that df 'uses' all the dose groups, not just those two being directly compared via (1). In this study, the degrees of freedom value was 32.

Development Of Power Computation Formulas

Standard statistical methodology would compare the computed value of the statistic t in (1) above to a critical value, T_{1-a}, from the Student's t distribution. The quantity a is the specified probability of falsely finding a (statistically) significant difference. If t is greater than T_{1-}

$_a$, then this is taken as evidence that there is an actual inhibition of ChE levels in the dose group compared with those in the placebo group. (Statistical significance would not, however, mean that the true inhibition is exactly the observed difference M_0-M_d.)

The power is the probability that the test will give a 'significant' result (i.e., $t>T_{1-a}$) when the actual inhibition in ChE is some value, D. That is:

$$(4) \qquad \text{Power} = P_D = \text{Prob}\{\, t > T_{1-a} \mid \text{actual inhibition is D}\,\}\ .$$

To find the power for any true inhibition level, it is necessary to know the distribution of t for different values of D. When D=0, t follows the well-known Student's t distribution. When D>0, however, t follows a non-central t distribution. This is more complicated than the simple t distribution and depends not only on degrees of freedom, df, but also on a non-centrality parameter, d. For this type of comparison, the non-centrality parameter is:

$$(5) \qquad d = D/\{s\,(1/n_0 + 1/n_d)^{½}\,\}$$

Here, s is the true inherent standard deviation (S is an estimator of s). The mathematical form of the non-central t distribution is rather complex but is incorporated within commercial power analysis software such as nQuery Advisor® (from Statistical Solutions). For spreadsheet-oriented power calculations, the non-central t distribution can be approximated quite well by the distribution of a Student's t variate plus the value d. That is,

$$(6) \qquad P_D = \text{Prob}\{\, t > T_{1-a} \mid D > 0\} \approx \text{Prob}\{\, t + d > T_{1-a} \mid D = 0\}$$

This means P_D can be found just by just computing d and finding:

$$(7) \qquad P_D = \text{Prob}\{\, t > T_{1-a} - d \mid D = 0\}$$

for the Student's t distribution using readily available tables or functions in spreadsheet software. Note from (7) that the power will increase as d increases. This means that power increases as the true ChE inhibition or the number of subjects increase. Conversely, the power decreases as the inherent variation, s, gets larger.

Computed Power For The Malathion Experiment

As seen in (5) above, some reasonable value for the inherent standard deviation, s, is needed in order to compute power. Using the calculations in (3), 10 different values of S, each an estimate of s, were obtained and listed in Table 2. Assuming that these are all representative of the same s, we selected the median of these 10 values (9.83% or, more simply, 10%) as a reasonable value to use in power calculations. Table 3 gives the power computed using the formulas above for D in the range of 10-30%. Here we also assume that a 5% 1-sided

significance level (i.e., a=0.05) would be used. As is evident from Table 3, the power to detect ChE inhibition levels is substantial down to D=15% for dose groups with 7 subjects and down to D=20% for dose groups with only 3 subjects. Even at a 15% inhibition level, the 0.73 power for 3-subject groups is not unreasonable.

We can also examine the power that would result over a wider range of between-subject variability, s. Table 4 shows power results computed using two additional values s=5 and s=13. These were chosen to bracket the S values in Table 2. Even for between-subject variation as high as 13%, the power for important inhibition levels (i.e. 20% or greater) is very good (over 0.70). For smaller values of s, the power remains excellent even down to 10% inhibition levels.

Power To Detect Trends

When a dose-response pattern exists, statistical methodology designed to look for trends or other such patterned effects should be more powerful than the more omnibus pairwise comparisons. A modification of the formulas above make it possible to determine the power for detecting linear trends in the dose response for this, and similar, experiments. The common test for trend utilizes an estimate, B, of the average change in ChE inhibition with dose (i.e., the 'slope'). B can be computed using linear contrasts in an analysis of variance or, equivalently, using the slope (but not the usual standard error) from a linear regression analysis. Regardless of how B is obtained, the relevant t-statistic used to test for trend is:

$$(8) \qquad t = B/SE_B$$

As before, this t statistic is be compared with a critical value from a Student's t distribution with df degrees of freedom. The estimated standard error of the slope, SE_B, is a function of the inherent variation estimate, S, (from formula 3 above), the spacing of the dose levels, and the number of subjects per dose:

$$(9) \qquad SE_B = S / \{ S_j n_i (d_i - D)^2 \}^{½} = S / Q$$

The value D is the weighted mean of all the dose levels in the experiment, i.e.,

$$(10) \qquad D = (S_j n_i d_i) / (S_j n_i)$$

It is important to note that this type of comparison actually tests for a *linear component* or *linear 'tendency'* for trend. It does not assume that the entire trend is linear. Thus, any upward change in ChE inhibition with dose should be potentially detectable with such a test.

The power associated with (8) is

$$(11) \quad \text{Power} = P_b = \text{Prob}\{ t > T_{1-a} \mid \text{actual trend is b} \}$$

In this case, t also follows a non-central t distribution, but with the non-centrality parameter, d, now being:

$$(12) \quad d = bQ/s$$

For power calculations, it usually more intuitive to express the slope, b, in terms of the inhibition at the highest dose, D_{max}, i.e.,

$$(13) \quad b = D_{max} / d_{max}$$

For example, because the largest dose in this study is d_{max}=15 mg/kg, an assumption that b=1 is equivalent to specifying that the inhibition at this dose is D_{max}=15%. Using the same assumptions as before, various powers to detect a linear trend in the malathion ChE human study were computed and are summarized in Table 5. As was the case with comparisons to placebo, the power to detect a linear trend in ChE inhibition is also excellent down to at least 15% (at the highest dose). The smaller the inherent between-subject variation, s, the greater the ability to detect even weaker trends. A comparison of the same levels of D in the pairwise tests (Table 4) with D_{max} in the trend tests (Table 5), shows that the power is, in fact, greater with the trend test. Thus, a test designed to detect a pattern in the data has more power than a 'general' test when that pattern actually occurs.

Table 1.
Configuration Of Dose Levels And Number Of Subjects Per Dose For The Malathion Study

Dose Level (mg/Kg)	Number of Subjects
0 (placebo)	11
0.5	3
1.5	3
5	7
10	7
15	7

Table 2.
Estimates Of Between-Subject Variability In Percent Decrease-From-Baseline Cholinesterase Levels

Time of Measurement after Start of Dosing	Estimated Between-Subject Standard Deviation, %	Degrees of Freedom
1 hour	7.28	32
2 hours	8.02	32
4 hours	11.74	32
8 hours	12.39	32
12 hours	12.09	32
1 day	8.12	32
2 days	10.15	32
4 days	10.55	30*
7 days	9.50	32
14 days	5.67	32
Minimum	***5.67***	
Maximum	***12.39***	
Median	***9.83***	

*2 missing measurements for this time period

Table 3.
Power To Detect True Percent ChE Inhibition[1] Using 1-Sided t Comparisons At A Significance Level Of 5% (a=0.05) And An Inherent Between-Subject Standard Deviation Of s=10%.

Actual ChE Inhibition (D),%	3 Subjects per Dose[2] (0.5 and 1.5 mg/kg)	7 Subjects per Dose[2] (5 10, and 15 mg/kg)
10	0.44	0.64
15	0.73	0.92
20	0.91	0.99
25	0.98	>0.99
30	>0.99	>0.99

[1] Inhibition = mean % deviation for placebo group – mean % deviation for dose group
[2] Placebo group had 11 subjects and degrees of freedom for S is equal to 32

Table 4.
Power To Detect True Percent ChE Inhibition[1] Using A Range Of Between-Subject Variation Levels. One-sided Pairwise t Comparisons At A Significance Level Of 5% (a=0.05) Were Assumed.

True ChE Inhibition (D), %	Subjects per dose[2]	Inherent Between-Subject Standard Deviation		
		s = 5	s = 10	s = 13
10	3	0.91	0.44	0.31
	7	0.99	0.64	0.46
15	3	>0.99	0.73	0.53
	7	>0.99	0.92	0.75
20	3	>0.99	0.91	0.75
	7	>0.99	0.99	0.93
25	3	>0.99	0.98	0.89
	7	>0.99	>0.99	0.99
30	3	>0.99	>0.99	0.96
	7	>0.99	>0.99	>0.99

[1] Inhibition = mean % deviation for placebo group – mean % deviation for dose group
[2] Placebo group had 11 subjects and the dose groups had either 3 subjects (0.5 and 1.5 mg/kg) or 7 subjects (5, 10, and 15 mg/kg). Degrees of freedom for S was equal to 32

Table 5.
Power To Detect A True Linear Trend In Percent ChE Inhibition[1] Using a Range of Between-Subject Variation Levels. A 1-Sided Test For Trend At A Significance Level Of 5% (a=0.05) Was Assumed.

True ChE Inhibition at the 15 mg/kg Dose, D_{max}, in the Presence of a Linear Trend with Dose	Resulting Power if Inherent Between-subject Standard Deviation is		
	s = 5	**s = 10**	**s = 13**
10 (b=0.667)	>0.99	0.74	0.54
15 (b=1.000)	>0.99	0.96	0.84
20 (b=1.333)	>0.99	>0.99	0.97
25 (b=1.667)	>0.99	>0.99	>0.99
30 (b=2.000)	>0.99	>0.99	>0.99

[1] Inhibition = mean % deviation for placebo group – mean % deviation for dose group
[2] Placebo group had 11 subjects and the dose groups had either 3 subjects (0.5 and 1.5 mg/kg) or 7 subjects (5, 10, and 15 mg/kg). Degrees of freedom for S was equal to 32